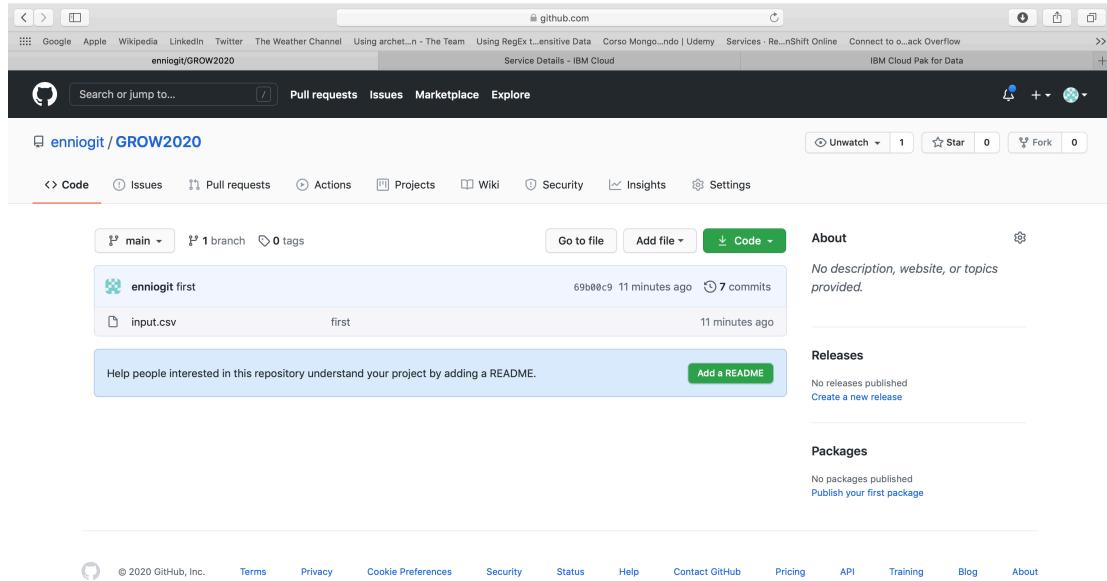
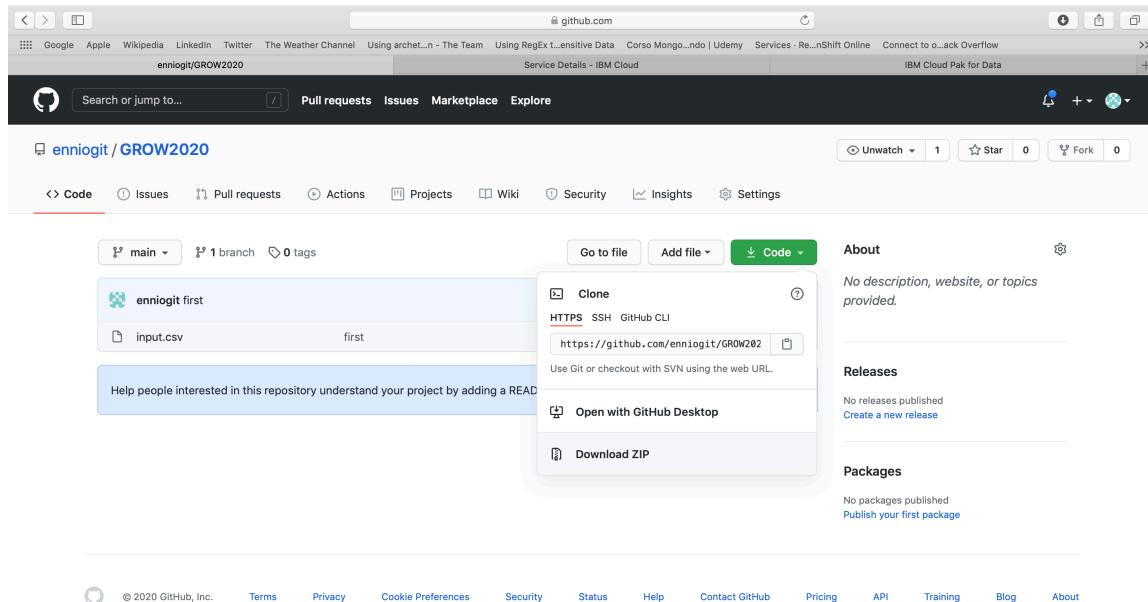


Go to

<https://github.com/enniogit/GROW2020>



Select Code and then download zip



After, unzip the file and have a look at the folder.

There will be some files included this document. We will see them later.

Go to <https://dataplatform.cloud.ibm.com/>

After login to be done by using the credential of registration you will get this screen

The screenshot shows the 'Create a project' interface. At the top, there's a header bar with various navigation links and a search bar. Below the header, there are two main options:

- Create an empty project**: This option is described as "Add the data you want to prepare, analyze, or model. Choose tools based on how you want to work: write code, create a flow on a graphical canvas, or automatically build models." It includes a "NEW" badge for "AutoAI experiment tool: Fully automated approach to building a classification or reg...". To the right, under "USE TO", are three items: "Prepare and visualize data", "Analyze data in notebooks", and "Train models".
- Create a project from a sample or file**: This option is described as "Get started fast by loading existing assets. Choose a project file from your system, or choose a curated sample project." To the right, under "USE TO", are three items: "Learn by example", "Build on existing work", and "Run tutorials".

At the bottom right of the screen is a blue speech bubble icon.

Select Create an empty project

The screenshot shows the 'New project' configuration screen. The top part is identical to the previous screenshot, showing the header and the two project creation options.

The main area is divided into several sections:

- Define project details**: Fields for "Name" (containing "Grow") and "Description" (containing "Project description").
- Define storage**: A section with two steps:
 - ① Select storage service **Add**: A note says "Add an object storage instance, and then return to this page and click Refresh."
 - ② Refresh
- Choose project options**: A checkbox "Restrict who can be a collaborator" is checked.

At the bottom right, there are "Cancel" and "Create" buttons.

And then insert the name of the project and select the Add link in the top right of the screen.

You'll get the following

The screenshot shows the IBM Cloud Pak for Data Services catalog interface. The URL is eu-de.dataplatform.cloud.ibm.com. The page title is "Cloud Object Storage". It displays a "Pricing plan" table comparing "Lite" and "Standard" plans. The "Lite" plan is free and includes 1 COS Service Instance, up to 25 GB/month storage, and up to 2,000 Class A requests per month. The "Standard" plan has no minimum fee and includes "There is no minimum fee, so you pay only for what you use." A "Create" button is visible on the right.

Plan	Features	Pricing
Lite	1 COS Service Instance Storage up to 25 GB/month Up to 2,000 Class A (PUT, COPY, POST, and LIST) requests per month Up to 20,000 Class B (GET and all others) requests per month Up to 10 GB/month of Data Retrieval Up to 5GB of egress (Public Outbound) Applies to aggregate total across all storage bucket classes	Free
Standard	There is no minimum fee, so you pay only for what you use.	See pricing details

Push the Create button.

After project creation You'll get the following screen.

The screenshot shows the IBM Cloud Pak for Data Project Overview screen for a project named "Grow". The URL is eu-de.dataplatform.cloud.ibm.com. The page title is "Service Details - IBM Cloud". The "Overview" tab is selected. Key statistics shown are 0 Assets and 1 Collaborator. The "Recent activity" section is empty. The "Collaborators" section lists "Ennio Picarelli" as an Admin. A "View all (1)" link is present. At the bottom, there is a "Readme" section with a note about Markdown syntax and a "Back to top" link.

Select Asset

Then via drag & drop upload the input.csv file and have a look at it.

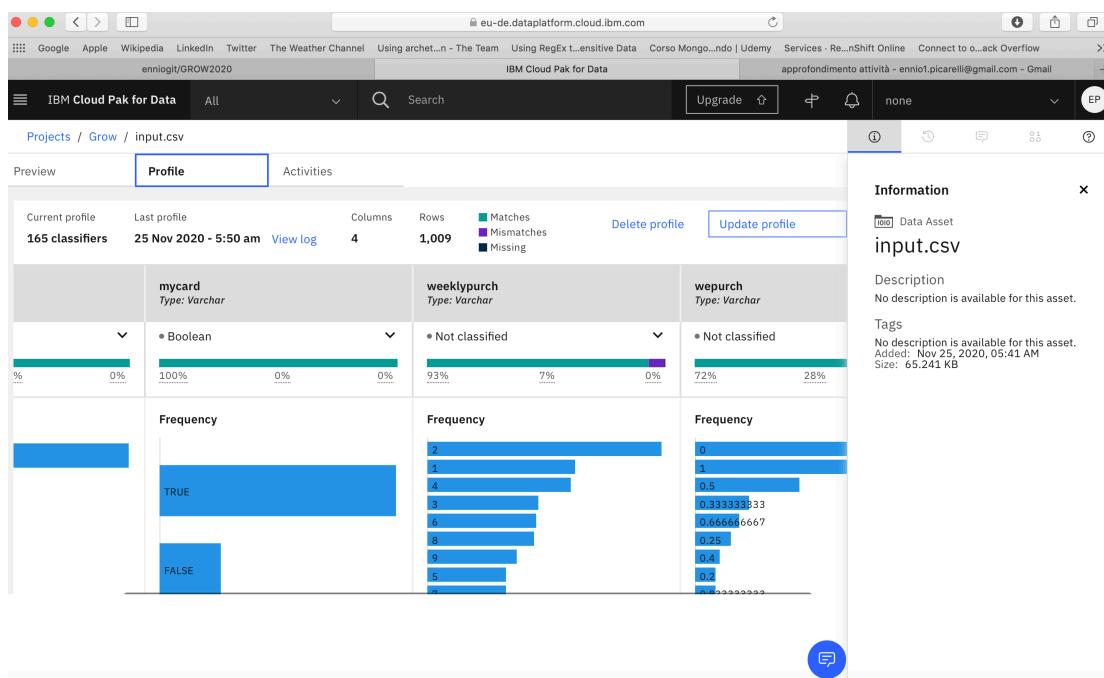
numcard	mycard	weeklypurch	wepurch
2	TRUE	14	0.928571429
2	TRUE	27	0.037037037
1	FALSE	102	0.019607843
2	TRUE	27	0.296296296
2	TRUE	8	0.625
2	TRUE	14	0
2	TRUE	8	0.25
2	TRUE	4	0
2	TRUE	3	1
2	TRUE	9	0.333333333
2	TRUE	4	0
2	TRUE	11	0.363636364
2	TRUE	3	1
2	TRUE	14	0.357142857
2	TRUE	13	0
-	-	-	-

In order to quickly analyze the content of the dataset, you can perform the Profile task (it will get a while).

The screenshot shows the 'Profile' tab selected in the top navigation bar. On the left, there's a message: 'Creating data profile' with a note: 'This process can take a while. Revisit the page at a later time or refresh it occasionally to check the status.' A blue 'Refresh' button is visible. On the right, the 'Information' panel displays the following details:

- Data Asset**: input.csv
- Description**: No description is available for this asset.
- Tags**: No description is available for this asset.
- Added**: Nov 25, 2020, 05:41 AM
- Size**: 18.374 KB

If you have a look at the profile results you can see that the `weeklypurch` has got some raw with null values and some other outliers (values below 0 or very high)



You can clean the dataset by using the refine function.
If you select Operation you can filter the raw corresponding to the missed values.

The screenshot shows the 'IBM Cloud Pak for Data' interface. On the left, there's a sidebar with 'Operation' selected. Below it, a 'Filter' section is open, showing a condition where 'weeklypurch' is not empty. The main area displays a table of data with 1009 rows. The right side shows the 'Information' panel with details like 'Data Source: input.csv', 'Convert column type: AUTOMATIC', and a 'Steps' section with one step. The 'DATA REFINERY FLOW NAME' is set to 'input.csv_flow'.

And you can replace with the 49 value (corresponding to the column mean) the negative or the values higher than 200.

This screenshot shows a similar interface but with a 'Calculate' operation. It uses the 'Is between two numbers' condition with values 0,100. A new column 'filtered' is created. The resulting data shows 1005 rows. The 'Information' panel shows a 'Steps' section with two steps: 'Data Source: input.csv' and 'Filter: JUST ADDED' (which is a result of the calculate operation). The 'DATA REFINERY FLOW NAME' is still 'input.csv_flow'.

After this you have created a data refinery flow

The screenshot shows the IBM Cloud Pak for Data interface for refining data. On the left, a table displays a portion of a CSV file with columns: numcard, mycard, weeklypurch, and wepurch. The table has 16 rows of data. In the center, a 'Refine data' panel shows the transformation steps:

- Data Source:** input.csv
- Convert column type:** AUTOMATIC (Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.)
- Remove empty rows:** Removed rows with blank or missing values in weeklypurch
- Conditional replace:** Replaced values for weeklypurch: weeklypurch where value is greater than 1000 as 49, weeklypurch where value is less than 0 as 49. Replaced all remaining values with weeklypurch.

On the right, the 'Information' panel shows details like the location (Grow), data refinery flow name (input.csv_flow), and data set name (input.csv_cleaned).

And you can save and create a job

The screenshot shows the IBM Cloud Pak for Data interface with a data refinery flow named 'input.csv_flow'. The 'Save and create a job' button is highlighted. The right-hand panel shows the flow's details, including its name, location (Grow), and output settings.

eu-de.dataplatform.cloud.ibm.com

GROW2020/input.csv at main · enniogit/GROW2020

Service Details - IBM Cloud

IBM Cloud Pak for Data

IBM Cloud Pak for Data

All

Search

Upgrade

none

EP

Projects / Grow / input.csv_flow

Create a job

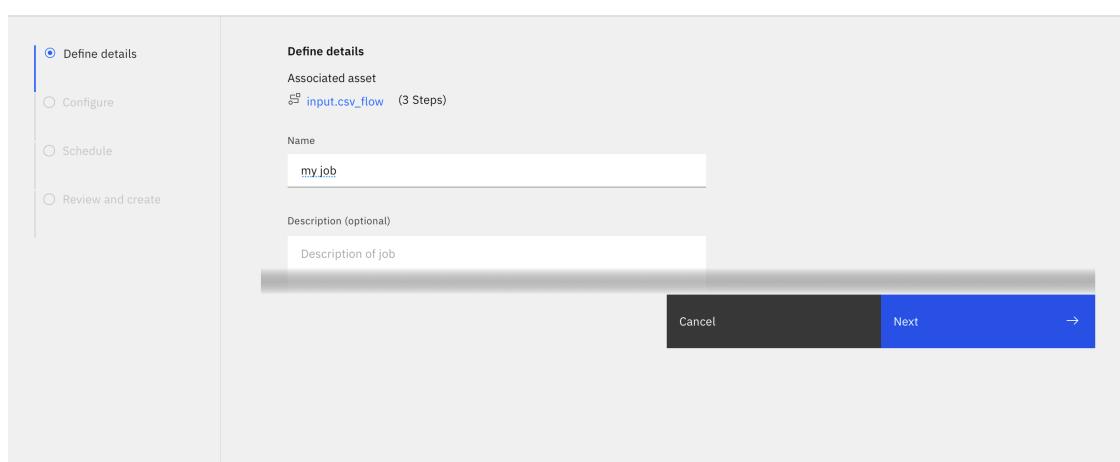
Define details

Associated asset
input.csv_flow (3 Steps)

Name
my job

Description (optional)
Description of job

Cancel Next →



eu-de.dataplatform.cloud.ibm.com

GROW2020/input.csv at main · enniogit/GROW2020

Service Details - IBM Cloud

IBM Cloud Pak for Data

IBM Cloud Pak for Data

All

Search

Upgrade

none

EP

Projects / Grow / input.csv_flow

Create a job

Review and create

Details

Associated asset
input.csv_flow

Name
my job

Description
Add Description

Configuration

Environment:
Default Data Refinery XS

Data assets

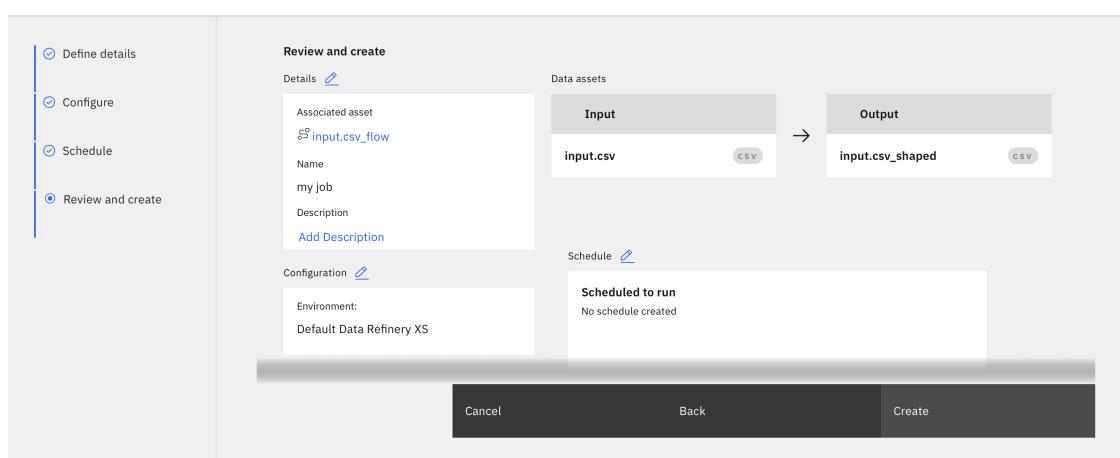
Input → Output

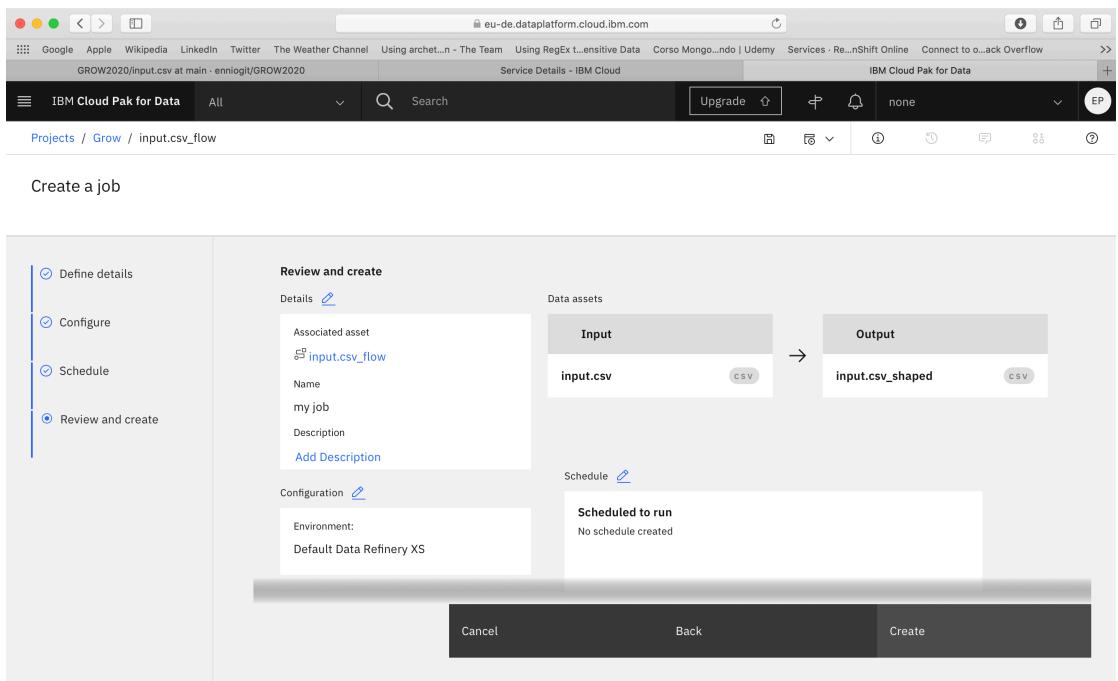
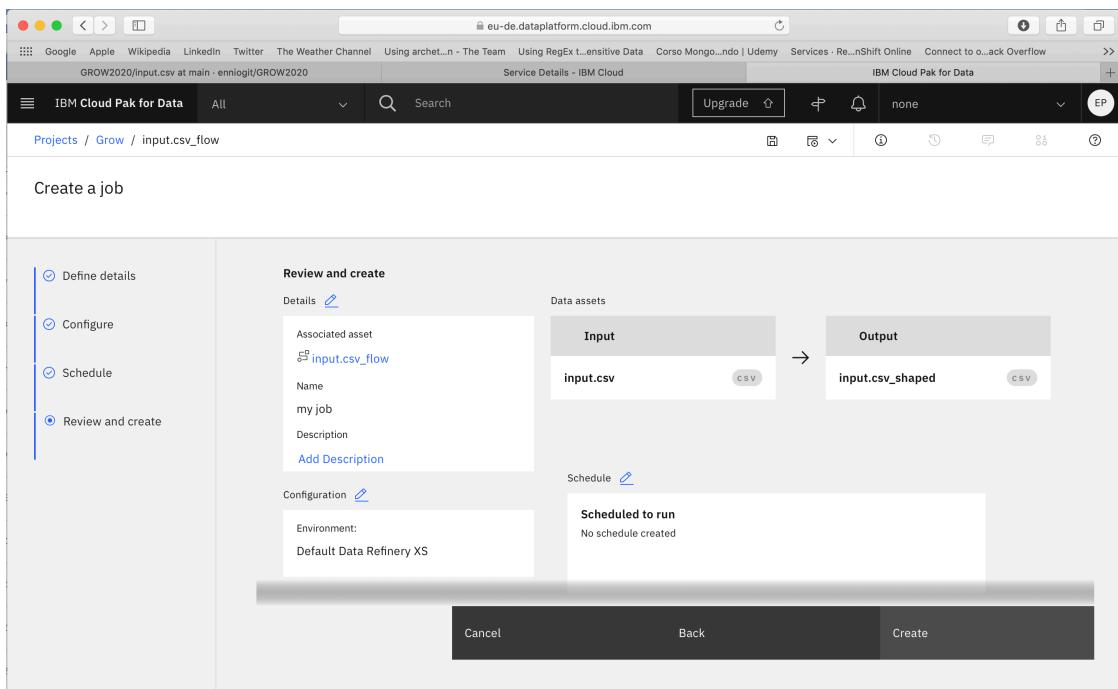
input.csv → input.csv_shaped

Schedule

Scheduled to run
No schedule created

Cancel Back Create





This job can then be executed immediately or scheduled

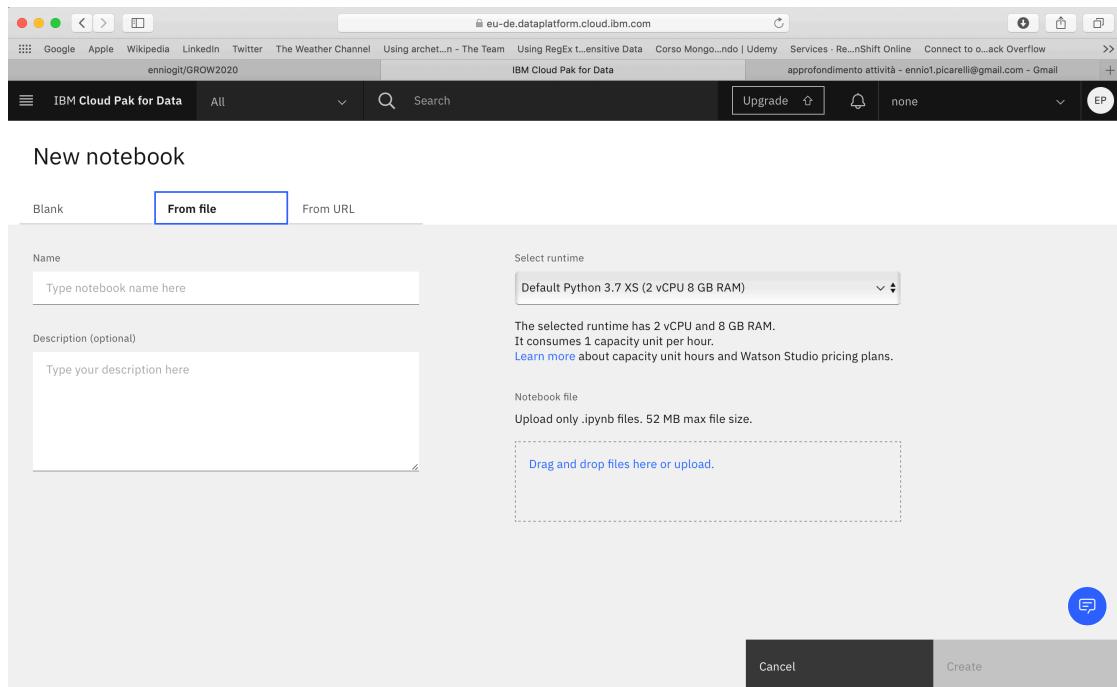
The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with links like Google, Apple, Wikipedia, LinkedIn, Twitter, The Weather Channel, Using archet...n - The Team, Using RegEx L...ensitive Data, Corso Mongo...ndo | Udemy, Services - Re...nShift Online, and Connect to o...ack Overflow. Below the navigation bar, the main title is "Service Details - IBM Cloud". The main content area shows a project named "my job" with a status message: "my job job run is successfully started. Click here to view the Run." Below this, there's a table titled "Runs (1)" with one entry: Start time (Nov 25, 2020 7:18:11 AM), Status (Running), Duration (---), Started by (Ennio Picarelli), and Action (three dots). A blue speech bubble icon is located at the bottom right of the main content area.

Now select add to project and from the popup Notebook

The screenshot shows the "Choose asset type" dialog box over the main IBM Cloud Pak for Data interface. The dialog has a title "Choose asset type" and a sub-section "Available asset types". It lists several options: Data, Notebook (highlighted with a black box and a tooltip: "Run small pieces of code to process your data and immediately view the results."), Connected data, AutoAI experiment, Notebook (NEW), Dashboard, Visual Recognition m..., MDM Configuration, Natural Language Cla..., Model from file, Deep learning experi..., Modeler flow, Metadata Import, Data Refinery flow, Streams flow, Decision Optimizatio..., and Data Refinery flows. To the right of the dialog, there's a "Data" panel with tabs for Load, Files, and Catalog, and a section for dropping files to upload. The main interface shows a "Projects / Growth" view with sections for Data assets and Data Refinery flows.

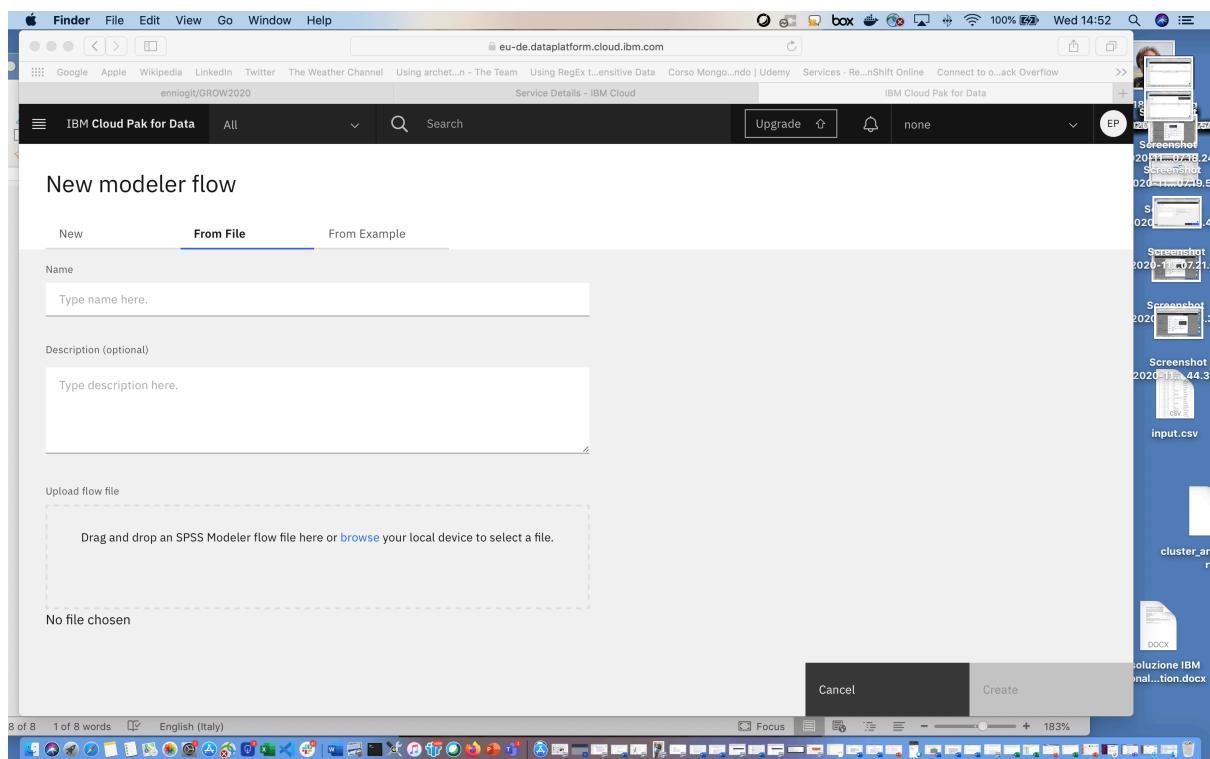
This is the environment that allows you to execute Jupyter notebooks with Python, R and Spark capabilities.

You can create a new Notebook and select the From file option then you can drag&drop the Datacleaning.ipynb file.



After the Notebook you can have a look as well at the modeler flow

The screenshot shows the 'Choose asset type' dialog in the IBM Cloud Pak for Data interface. The dialog lists various asset types: Data, Connection, Connected data, AutoAI experiment, Notebook (highlighted with a blue border and labeled 'NEW'), Visual Recognition m..., MDM Configuration, Model from file, Deep learning experi..., Modeler flow, Metadata Import, Data Refinery flow, Streams flow, and Decision Optimizatio... (all labeled 'NEW'). The 'Modeler flow' option is highlighted with a black box and a tooltip: 'Create an SPSS Modeler flow to prepare or shape data, train or deploy a model, or transform data and export it back to a database table or a file.' The background shows the main interface with sections for 'Projects / Grow' and 'Assets' (selected). The 'Assets' section shows 'Data assets' (input.csv_shaped, input.csv) and 'Notebooks' (Datacleaning, prova).



Drag & drop Datacleaning.str and then select create.
Repeat the same operation with cluster_analysis